What is claimed is:

1. A method for constructing a variant set for a biopolymer of interest, the method comprising:

a) identifying, using a plurality of rules, a plurality of positions in said biopolymer of interest and, for each respective position in said plurality of positions, one or more substitutions for the respective position, wherein the plurality of positions and the one or more substitutions for each respective position in the plurality of positions collectively define a biopolymer sequence space;

b) selecting a variant set, wherein said variant set comprises a plurality of variants of said biopolymer of interest and wherein said variant set is a subset of said biopolymer sequence space;

c) measuring a property of all or a portion of the variants in said variant set;

d) modeling a sequence-activity relationship between (i) one or more substitutions at one or more positions of the biopolymer of interest represented by the variant set and (ii) the property measured for all or said portion of the variants in the variant set; and

e) redefining said variant set to comprise variants that include substitutions in said plurality of positions that are selected based on a function of said sequence-activity relationship.

2. The method of claim 1, the method further comprising repeating said measuring, modeling, and, optionally, said redefining, until a variant in said variant set exhibits a value for said property that exceeds a predetermined value.

3. The method of claim 2 wherein said predetermined value is a value that is greater than the value for the property that is exhibited by said biopolymer of interest.

4. The method of claim 1, the method further comprising repeating said measuring, modeling, and, optionally, said redefining, until a variant in said variant set exhibits a value for said property that is less than a predetermined value.

5. The method of claim 4 wherein said predetermined value is a value that is less than the value for the property that is exhibited by said biopolymer of interest.

6. The method of claim 1, the method further comprising repeating said measuring, modeling, and, optionally, said redefining, a predetermined number of times.

7. The method of claim 6 wherein said predetermined number of times is two, three, four, or five.

8. The method of claim 1 wherein said sequence-activity relationship comprises a plurality of values and wherein each value in said plurality of values describes a relationship between (i) a substitution at a position in said plurality of positions represented by said all or said portion of the variants in said variant set and said property, (ii) a plurality of substitutions at a position in said plurality of positions represented by said all or said portion of the variants in said variant set and said property, or (iii) one or more substitutions in one or more positions in said plurality of positions represented by said all or said portion of the variants in said variant set and said property.

9. The method of claim 8 wherein said modeling comprises regressing:

$$V_{measured} = W_{11}P_1S_1 + W_{12}P_1S_2 + \quad + W_{1N}P_1S_N + \quad + W_{M1}P_MS_1 + W_{M2}P_MS_2 + \quad + W_{MN}P_MS_N$$

wherein,

$V_{measured}$ represents the property measured in variants in said variant set;

W_{MN} = is a value in said plurality of values;

P_M = is a position in said biopolymer of interest in said plurality of positions in said biopolymer of interest; and

S_N = is a substitution in the one or more positions for a position in the plurality of positions in said biopolymer of interest.

10. The method of claim 9 wherein said regressing comprises linear regression, non-linear regression, logistic regression, multivariate data analysis, or partial least squares projection to latent variables.

11. The method of claim 1 wherein said modeling comprises computation of a neural network, computation of a bayesian model, a generalized additive model, a support vector machine, or classification using a regression tree.

12. The method of claim 1 wherein said modeling comprises boosting or adaptive boosting.

13. The method of claim 1 wherein said redefining further comprises:

computing a predicted score for a population of variants of said biopolymer of interest using said sequence-activity relationship, wherein each variant in said population of variants includes a substitution at one or more positions in said plurality of positions in said biopolymer of interest; and

selecting said variant set from among said population of variants as a function of the predicted score received by each variant in said set of variants.

14. The method of claim 13, the method further comprising

ranking said population of variants, wherein each variant in said population of variants is ranked based on the predicted score received by the variant based upon the sequence-activity relationship; and

said selecting comprising accepting a predetermined percentage of the top ranked variants in said population of variants for said variant set.

15. The method of claim 13, wherein a respective variant in said population of variants is selected for said variant set when the predicted score of the respective variant exceeds a predetermined value.

16. The method of claim 1 wherein said redefining step (e) further comprises redefining said variant set to comprise one or more variants each having a substitution in a position in said plurality of positions not present in any variant in the variant set selected by said selecting step (b).

17. The method of claim 1 wherein

said modeling a sequence-activity relationship (d) further comprises modeling a plurality of sequence-activity relationships, wherein each respective sequence-

activity relationship in said plurality of sequence-activity relationships describes the relationship between (i) one or more substitutions at one or more positions of the biopolymer of interest represented by the variant set and (ii) the property measured for all or said portion of the variants in the variant set; and

said redefining said variant set (e) comprises redefining said variant set to comprise variants that include substitutions in said plurality of positions that are selected based on a combination of said plurality of sequence-activity relationships.

18. The method of claim 17, the method further comprising:

repeating said measuring based upon said redefined variant set, wherein a property of all or a portion of the variants in the redefined variant set is measured; and

weighting each respective sequence-activity relationship in said plurality of sequence activity relationships based on an agreement between (i) measured values for the property of variants in said redefined variant set and (ii) values for the property of variants in said redefined variant set that were predicted by said respective sequence-activity relationship, wherein

a first sequence-activity relationship that achieves better agreement between measured and predicted values than a second sequence-activity relationship receives a higher weight than said second sequence-activity relationship.

19. The method of claim 17 wherein said redefining step (e) further comprises redefining said variant set to comprise one or more variants each having a substitution in a position in said plurality of positions not present in any variant in the variant set selected by said selecting step (b).

20. The method of claim 18 wherein said redefining step (e) further comprises redefining said variant set to comprise one or more variants each having a substitution in a position in said plurality of positions not present in any variant in the variant set selected by said selecting step (b).

21. The method of claim 1, wherein

the contribution of each respective rule in said plurality of rules to said biopolymer sequence space is independently weighted by a rule weight in a plurality of rule weights corresponding to the respective rule; and

the method further comprising, prior to said redefining step (e), the steps of:

adjusting one or more rule weights in said plurality of rule weights based on a comparison, for each respective variant in the variant set, (i) a value assigned to the respective variant by said sequence-activity relationship, and (ii) a score assigned by the plurality of rules to the respective variant; and

repeating said identifying step using said rule weights, thereby redefining said plurality of positions and, for each respective position in said plurality of positions, redefining the one or more substitutions for the respective position; and

wherein

said redefining step (e) further comprises redefining said variant set to comprise one or more variants each having a substitution in a position in said redefined plurality of positions not present in any variant in the variant set selected by said initial selecting step (b).

22. The method of claim 21 wherein

said modeling a sequence-activity relationship (d) further comprises modeling a plurality of sequence-activity relationships, wherein each respective sequence-activity relationship in said plurality of sequence-activity relationships describes the relationship between (i) one or more substitutions at one or more positions of the biopolymer of interest represented by the variant set and (ii) the property measured for all or said portion of the variants in the variant set; and

said redefining said variant set (e) comprises redefining said variant set to comprise variants that include substitutions in said plurality of positions that are selected based on a combination function of said plurality of sequence-activity relationships.

23. The method of claim 22, the method further comprising:

repeating said measuring based upon said redefined variant set, wherein a property of all or a portion of the variants in the redefined variant set is measured; and

weighting each respective sequence-activity relationship in said plurality of sequence activity relationships based on an agreement between (i) measured values for the property of variants in said redefined variant set and (ii) values for the property of variants in said redefined variant set that were predicted by said respective sequence-activity relationship, wherein

a first sequence-activity relationship that achieves better agreement between measured and predicted values than a second sequence-activity relationship receives a higher weight than said second sequence-activity relationship.

24. The method of claim 1 wherein said biopolymer of interest is a polypeptide, a polynucleotide, a small inhibitory RNA molecule (siRNA), or a polyketide.

25. The method of claim 1 wherein said biopolymer of interest is a protein kinase, a protein phosphatase, a protease, a receptor, a G-protein coupled receptor, a cytokine, a growth factor or an antigen from an infectious pathogen.

26. The method of claim 1 wherein said biopolymer of interest is a cytochrome P450, a lipase, an esterase, a peptidase, a transferase, a polymerase, or a depolymerase.

27. The method of claim 1 wherein said plurality of positions comprises five or more positions.

28. The method of claim 1 wherein said plurality of positions comprises ten or more positions.

29. The method of claim 1 wherein said plurality of rules comprises two or more rules.

30. The method of claim 1 wherein said plurality of rules comprises five or more rules.

31. The method of claim 1 wherein, a rule in said plurality of rules assigns a score to a variant of said biopolymer of interest by considering a lineup of a plurality of sequences that are homologous to said biopolymer of interest.

32. The method of claim 1 wherein, a rule in said plurality of rules assigns a score to a variant of said biopolymer of interest by considering structural variations in one or more three dimensional structures of biopolymers that are homologous to said biopolymer of interest.

33. The method of claim 1 wherein a rule in said plurality of rules assigns a score to a variant of said biopolymer of interest by considering variations in a substitution matrix for said biopolymer of interest.

34. The method of claim 31 wherein said substitution matrix is a universal substitution matrix.

35. The method of claim 1 wherein a first rule in said plurality of rules assigns a score to a variant of said biopolymer of interest based upon a binding pocket analysis of a structural model of said biopolymer of interest or a structural model of a homolog of said biopolymer of interest.

36. The method of claim 1, wherein said identifying combines a score from each rule in said plurality of rules for a variant of a biopolymer of interest.

37. The method of claim 36 wherein said combining comprises adding (i) a first score from a first rule in said plurality rules and (ii) a second score from a second rule in said plurality rules for said variant of a biopolymer of interest.

38. The method of claim 36 wherein said combining comprises multiplying (i) a first score from a first rule in said plurality rules and (ii) a second score from a second rule in said plurality rules for said variant of a biopolymer of interest.

39. The method of claim 1 wherein said variant set consists of between 5 and 200 variants of said biopolymer of interest.

40. The method of claim 1 wherein said variant set consists of between 15 and 50 variants of said biopolymer of interest.

41. The method of claim 1 wherein said selecting said variant set (b) comprises applying a monte carlo algorithm, a genetic algorithm, or a combination thereof, to construct said variant set, with the provisos that:

(i) each variant in all or portion of said variant set has a number of substitutions that is between a first value and a second value; and

(ii) a number of different pairs of substitutions collectively represented by said variant set is above a predetermined number.

42. The method of claim 41 wherein said first value is two substitutions and said second value is twenty substitutions.

43. The method of claim 41 wherein said first value is four substitutions and said second value is ten substitutions.

44. The method of claim 41 wherein said predetermined number is thirty.

45. The method of claim 41 wherein said predetermined number is one hundred.

46. The method of claim 1 wherein said selecting said variant set (b) comprises:

dividing said biopolymer of interest into one or more functional domains; and for each respective functional domain in said one or more functional domains,

applying a monte carlo algorithm, a genetic algorithm, or a combination thereof, in order to identify substitutions at positions in the plurality of positions that are in said respective functional domain for inclusion in one or more variants in said variant set, with the provisos that:

all or a portion of the variants in the variant set contains a predetermined number of substitutions at positions from each of the one or more functional domains; and

a number of different pairs of substitutions at positions in each of the one or more functional domains that is collectively represented by the variant set is above a threshold value.

47. The method of claim 46 wherein said predetermined number is between two and twenty.

48. The method of claim 46 wherein said predetermined number is between four and ten.

49. The method of claim 46 wherein said threshold value is thirty.

50. The method of claim 46 wherein said threshold value is one hundred.

51. The method of claim 1 wherein

said measuring comprises synthesizing all or said portion of the variants in said variant set, and wherein

said property of a variant in said variant set is an antigenicity of said variant, an immunogenicity of said variant, an immunomodulatory activity of said variant, a catalysis of a chemical reaction by said variant, a thermostability of said variant, a level of expression of said variant in a host cell, a susceptibility of said variant to a post-translational modification, a killing of pathogenic organisms or viruses resulting from activity of said variant or a modulation of a signaling pathway by said variant.

52. The method of claim 1 wherein said sequence-activity relationship has the form:

$$Y = f(w_1x_1, w_2x_2, \quad w_ix_i)$$

wherein,

Y is a quantitative measure of said property;

x_i is a descriptor of a substitution, a combination of substitutions, or a component of one or more substitutions, at one or more positions in said plurality of positions;

w_i is a weight applied to descriptor x_i; and

f( ) is a mathematical function.

53. The method of claim 52 wherein said modeling comprises regressing:

$$Y = f(w_1x_1, w_2x_2, \quad w_ix_i).$$

54. The method of claim 53 wherein regressing comprises linear regression, non-linear regression, logistic regressing, or partial least squares projection to latent variables.

55. The variant of claim 2 that exhibits a value for said property that exceeds a predetermined value.

56. The variant of claim 55 wherein said predetermined value is a value that is greater than the value for the property that is exhibited by said biopolymer of interest.

57. A composition comprising the variant of claim 55, and a carrier.

58. A plurality of nucleic acid sequences comprising nucleotide sequences encoding all or a portion of the variants in the redefined variant set of step e) of claim 1.

59. All or a portion of the variants in the redefined variant set of step e) of claim 1.

60. A plurality of nucleic acid sequences comprising nucleotide sequences encoding variants in the redefined variant set of an instance of step e) of claim 2.

61. All or a portion of the variants in the redefined variant set of an instance of step e) of claim 2.

62. A plurality of nucleic acid sequences comprising nucleotide sequences encoding variants in the redefined variant set of an instance of step e) of claim 4.

63. The variants in the redefined variant set of an instance of step e) of claim 4.

64. A plurality of nucleic acid sequences comprising nucleotide sequences encoding variants in the redefined variant set of an instance of step e) of claim 6.

65. The variants in the redefined variant set of an instance of step e) of claim 6.

66. A population of cells comprising nucleic acid sequences encoding a plurality of variants in the redefined variant set of step e) of claim 1.

67. A population of cells comprising the variants in the redefined variant set of step e) of claim 1.

68. A population of cells comprising nucleic acid sequences encoding variants in the redefined variant set of an instance of step e) of claim 2.

69. A population of cells comprising variants in the redefined variant set of an instance of step e) of claim 2.

70. A population of cells comprising nucleotide sequences encoding variants in the redefined variant set of an instance of step e) of claim 4.

71. A population of cells comprising nucleotide sequences encoding variants in the redefined variant set of an instance of step e) of claim 4.

72. A population of cells comprising nucleotide sequences encoding variants in the redefined variant set of an instance of step e) of claim 6.

73. A population of cells comprising variants in the redefined variant set of an instance of step e) of claim 6.

74. A method of weighting a plurality of selection rules for use in selecting a plurality of positions in a biopolymer of interest and, for each respective position in said plurality of positions, one or more substitutions for the respective positions, the method comprising:

a) identifying, using said plurality of selection rules, a plurality of positions in a biopolymer of interest and, for each respective position in said plurality of positions, one or more substitutions for the respective position, wherein

the plurality of positions and the one or more substitutions for each respective position in the plurality of positions collectively define a biopolymer sequence space, and

the contribution of each respective rule in said plurality of rules to said biopolymer sequence space is independently weighted by a rule weight in a plurality of rule weights corresponding to the respective rule;

b) selecting a variant set, wherein said variant set comprises a plurality of variants of said biopolymer of interest and wherein said variant set is a subset of said biopolymer sequence space;

c) measuring a property of all or a portion of the variants in said variant set;

d) modeling a sequence-activity relationship between (i) one or more substitutions at one or more positions of the biopolymer of interest represented by the variant set and (ii) the property measured for all or said portion of the variants in the variant set; and

e) adjusting one or more rule weights in said plurality of rule weights based on a comparison, for each respective variant in the variant set, (i) a value assigned to the respective variant by the sequence-activity relationship, and (ii) a score assigned by the plurality of rules to the respective variant;

f) repeating said identifying, selecting, measuring, modeling, and adjusting for each biopolymer of interest in a plurality of biopolymers of interest.

75. The method of claim 74, the method further comprising, prior to said repeating (f):

(i) modeling a sequence-activity relationship between (a) one or more substitutions at one or more positions of the biopolymer of interest represented by the variant set and (b) the property measured for all or said portion of the variants in the variants in the variant set; and

(ii) redefining said variant set to comprise variants that include substitutions in said plurality of positions that are selected based on a function of said sequence-activity relationship.

76. The method of claim 75, wherein

said modeling a sequence-activity relationship further comprises modeling a plurality of sequence-activity relationships, wherein each respective sequence-activity relationship in said plurality of sequence-activity relationships describes the relationship between (i) one or more substitutions at one or more positions of the biopolymer of interest represented by the variant set and (ii) the property measured for all or said portion of the variants in the variant set; and

said redefining said variant set comprises redefining said variant set to comprise variants that include substitutions in said plurality of positions that are selected based on a combination function of said plurality of sequence-activity relationships.

77. The method of claim 76, the method further comprising:

repeating said measuring (c) using said redefined variant set, wherein a property of all or a portion of the variants in the redefined variant set is measured; and

weighting each respective sequence-activity relationship in said plurality of sequence activity relationships based on an agreement between (i) measured values for the property of variants in said redefined variant set and (ii) values for the property of variants in said redefined variant set that were predicted by said respective sequence-activity relationship, wherein

a first sequence-activity relationship that achieves better agreement between measured and predicted values than a second sequence-activity relationship receives a higher weight than said second sequence-activity relationship.

78. The method of claim 77 wherein said redefining further comprises redefining said variant set to comprise one or more variants each having a substitution in a position in said plurality of positions not present in any variant in the variant set selected by said selecting step (b).

79. The method of claim 74 wherein said plurality of biopolymers of interest represent a biopolymer class.

80. The method of claim 79 wherein said biopolymer class is protein, deoxyribose nucleic acid (DNA), ribsose nucleic acid (RNA), or polyketide.

81. The method of claim 74 wherein said plurality of biopolymers of interest represent a biopolymer subclass.

82. The method of claim 81 wherein said biopolymer subclass is protein kinases, protein phosphatases, proteases, receptor proteins, cytokines, growth factors, thrombomodulatory molecules, integrins, antigens from infectious pathogens, cytochrome P450s, lipases, esterases, peptidases, transferases, polymerases, depolymerases, type II polyketides, type I polyketides, non-ribosomal peptides or terpenes.

83. The method of claim 74 wherein said plurality of biopolymers of interest represent a biopolymer class and wherein said property is an antigenicity of said variant, an immunogenicity of said variant, an immunomodulatory activity of said variant, a catalysis of a chemical reaction by said variant, a thermostability of said variant, a level of expression of said variant in a host cell, a susceptibility of said variant to a post-translational modification, a killing of pathogenic a organism or a virus resulting from activity of said variant, or a modulation of a signaling pathway by said variant.

84. The method of claim 75, the method further comprising repeating said measuring (c), modeling (d), adjusting (e), modeling (i), and redefining (ii) until a variant in said variant set exhibits a value for said property that exceeds a predetermined value.

85. The method of claim 84 wherein said predetermined value is a value that is greater than the value for the property that is exhibited by said biopolymer of interest.

86. The method of claim 75, the method further comprising repeating said measuring (c), modeling (d), adjusting (e), modeling (i), and redefining (ii) until a variant in said variant set exhibits a value for said property that is less than a predetermined value.

87. The method of claim 86 wherein said predetermined value is a value that is less than the value for the property that is exhibited by said biopolymer of interest.

88. The method of claim 75, the method further comprising repeating said measuring (c), modeling (d), adjusting (e), modeling (i), and redefining (ii) a predetermined number of times.

89. The method of claim 88 wherein said predetermined number of times is two, three, four, or five.

90. A plurality of nucleic acid sequences comprising nucleotide sequences encoding all or a portion of the variants in the redefined variant set of step (ii) of claim 75.

91. All or a portion of the variants in the redefined variant set of step (ii) of claim 75.

92. The variant of claim 84.

93. A nucleic acid encoding the variant of claim 84.

94. The variant of claim 86.

95. A nucleic acid encoding the variant of claim 86.

96. A plurality of nucleic acid sequences comprising nucleotide sequences encoding all or a portion of the variants in the redefined variant set of an instance of step (ii) of claim 88.

97. All or a portion of the variants in the redefined variant set of an instance of step (ii) of claim 88.

98. A computer program product for use in conjunction with a computer system, the computer program product comprising a computer readable storage medium and a computer program mechanism embedded therein, the computer program mechanism comprising:

a knowledge base comprising a plurality of rules; and

an expert module for constructing a variant set for a biopolymer of interest, the expert module comprising:

instructions for identifying, using said plurality of rules, a plurality of positions in said biopolymer of interest and, for each respective position in said plurality of positions, one or more substitutions for the respective position, wherein the plurality of positions and the one or more substitutions for each respective position in the plurality of positions collectively define a biopolymer sequence space;

instructions for selecting a variant set, wherein said variant set comprises a plurality of variants of said biopolymer of interest and wherein said variant set is a subset of said biopolymer sequence space;

instructions for measuring or receiving a measurement a property of all or a portion of the variants in said variant set;

instructions for modeling a sequence-activity relationship between (i) one or more substitutions at one or more positions of the biopolymer of interest represented by the variant set and (ii) the property measured for all or said portion of the variants in the variant set; and

instructions for redefining said variant set to comprise variants that include substitutions in said plurality of positions that are selected based on a function of said sequence-activity relationship.

99. The computer program product of claim 98 wherein said knowledge base further stores a substitution matrix that is referenced by a rule in said plurality of rules during an instance of said instructions for identifying.

100. The computer program product of claim 98 wherein said knowledge base further stores a conservation index that is referenced by a rule in said plurality of rules during an instance of said instructions for identifying.

101. The computer program product of claim 98 wherein said knowledge base further stores mutation effect data that is referenced by a rule in said plurality of rules during an instance of said instructions for identifying.

102. The computer program product of claim 98 wherein said knowledge base further stores three dimensional structural information for said biopolymer of interest and/or a homolog of said biopolymer of interest that is referenced by a rule in said plurality of rules during an instance of said instructions for identifying.

103. The computer program product of claim 98, the expert module further comprising instructions for repeating said instructions for measuring, instructions for modeling, and, optionally, said instructions for redefining, until a variant in said variant set exhibits a value for said property that exceeds a predetermined value.

104. The computer program product of claim 98, the expert module further comprising instructions for repeating said instructions for measuring, instructions for modeling, and, optionally, said instructions for redefining, until a variant in said variant set exhibits a value for said property that is less than a predetermined value.

105. The computer program product of claim 98, the expert module further comprising instructions for repeating said instructions for measuring, instructions for modeling, and, optionally, said instructions for redefining, a predetermined number of times.

106. The computer program product of claim 98, wherein the instructions for selecting said variant set comprise instructions for applying a monte carlo algorithm, a genetic algorithm, or a combination thereof, to construct said variant set, with the provisos that:

(i) each variant in all or portion of said variant set has a number of substitutions that is between a first value and a second value; and

(ii) a number of different pairs of substitutions collectively represented by said variant set is above a predetermined number.

107. The computer program product of claim 106 wherein said first value is two substitutions and said second value is twenty substitutions.

108. The computer program product of claim 106 wherein said first value is four substitutions and said second value is ten substitutions.

109. The computer program product of claim 106 wherein said predetermined number is thirty.

110. The computer program product of claim 106 wherein said predetermined number is one hundred.

111. The computer program product of claim 98 wherein said instructions for selecting the variant set comprise instructions for dividing said biopolymer of interest into one or more functional domains, and for each respective functional domain in said one or more functional domains, instructions for applying a monte carlo algorithm, a genetic algorithm, or a combination thereof, in order to identify substitutions at positions in the plurality of positions that are in said respective

functional domain for inclusion in one or more variants in said variant set, with the provisos that:

all or a portion of the variants in the variant set contains a predetermined number of substitutions at positions from each of the one or more functional domains; and

a number of different pairs of substitutions at positions in each of the one or more functional domains that is collectively represented by the variant set is above a threshold value.

112. The method of claim 111 wherein said predetermined number is between two and twenty.

113. The method of claim 111 wherein said predetermined number is between four and ten.

114. The method of claim 111 wherein said threshold value is thirty.

115. The method of claim 111 wherein said threshold value is one hundred.

116. A computer system comprising:

a central processing unit;

a memory, coupled to the central processing unit, the memory storing a knowledge base and an expert module, wherein

the knowledge base comprises a plurality of rules; and

the expert module is for constructing a variant set for a biopolymer of interest and comprises:

instructions for identifying, using said plurality of rules, a plurality of positions in said biopolymer of interest and, for each respective position in said plurality of positions, one or more substitutions for the respective position, wherein the plurality of positions and the one or more substitutions for each respective position in the plurality of positions collectively define a biopolymer sequence space;

instructions for selecting a variant set, wherein said variant set comprises a plurality of variants of said biopolymer of interest and wherein said variant set is a subset of said biopolymer sequence space;

instructions for measuring or receiving a measurement a property of all or a portion of the variants in said variant set;

instructions for modeling a sequence-activity relationship between (i) one or more substitutions at one or more positions of the biopolymer of interest represented by the variant set and (ii) the property measured for all or said portion of the variants in the variant set; and

instructions for redefining said variant set to comprise variants that include substitutions in said plurality of positions that are selected based on a function of said sequence-activity relationship.